Erciyes Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 35, Sayı 3, 2019

Erciyes University
Journal of Institue Of Science and Technology
Volume 35, Issue 3, 2019

Genetik Algoritma Kullanılarak Verilerin Karma Normal Modele Dayalı Kümelenmesi

Maruf Gögebakan*[1], Tayfun Servi[2]

*[1] Bandırma Onyedi Eylül Üniversitesi Denizcilik Fakültesi Denizcilik İşletmeleri Yönetimi Bölümü, BANDIRMA / BALIKESİR

[2] Adıyaman Üniversitesi İktisadi ve İdari Bilimler Fakültesi İktisat Bölümü, ADIYAMAN

(Alınış / Received: 11.11.2019, Kabul / Accepted: 16.12.2019, Online Yayınlanma / Published Online: 31.12.2019)

Anahtar Kelimeler
Genetik Algoritma,
Karma Normal Model,
Modele dayalı kümeleme,
Bilgi kriterleri

Öz: Bu çalışmada, çok değişkenli homojen ve heterojen büyük verilerin kümelemesi için yeni bir kümeleme algoritması geliştirildi. Heterojen verideki parçalanmalar, kümelerin sayısını ve yerini belirler. Heterojen verilerdeki parçalanmaların sayısı, hem grafiksel hem de hesaplamalı yöntemlere dayalı olarak belirlenir. Grafiksel yöntemlerde her bir değişkenin olasılık grafikleri, hesaplamalı yöntemlerde ise her değişkenin tek değişkenli karma normal dağılımları kullanılır. Genetik algoritmalar, heterojen verideki parçalanmalara karşılık gelen kümelenme merkezlerinin yerini ve yapısını belirlemede kullanılır. Kümelenme merkezlerinin sayısı ve yapısına dayalı belirlenen modeller Karma normal dağılımlar kullanılarak elde edilir. Karma normal modellerdeki her bir küme merkezi, değişkenlerdeki parçalanmalara karşılık gelir. Karma normal modeller arasından veri yapısına uyan en iyi karma model karma normal dağılımlardan elde edilen bilgi kriterleri kullanılarak elde edilir.

Normal Mixture Model-Based Clustering of Data Using Genetic Algorithm

Keywords
Genetic Algorithm,
Gaussian Mixture Models,
Model Based Clustering,
Information Criteria

Abstract: In this study, a new clustering algorithm was developed for the clustering of multivariate homogeneous and heterogeneous big data. Fragments in heterogeneous data determine the number and location of clusters. The number of fragments in heterogeneous data is determined based on both graphical and computational methods. In graphical methods, the probability graphs of each variable are used, while the computational methods use the univariate mixture normal distributions of each variable. Genetic algorithms are used to determine the location and structure of clustering centers corresponding to fragmentation in heterogeneous data. Determined models based on the number and structure of cluster centers is obtained by using mixture normal distributions. Each cluster center in mixture normal models corresponds to fragmentation in the variables. The best mixture model that matches the data structure from the mixture normal models is obtained by using the information criteria obtained from mixture normal distributions.

Denizcilik İşletmeleri Yönetimi Bölümü, Bandırma Onyedi Eylül Üniversitesi, Balıkesir
İktisat Bölümü, Adıyaman Üniversitesi, Adıyaman
*İlgili yazar, email: mgogebakan@bandirma.edu.tr

1. Giriş

Sonlu karma dağılımlarda modele dayalı kümeleme, p-boyutlu çok değişkenli veriyi altgruplara ayırmak için kullanılan en etkili kümeleme yöntemlerindendir [1]. Çok değişkenli normal dağılımların karmasındaki her bileşeni, çok değişkenli heterojen verideki bir kümeye karşılık gelir[2]. Çok değişkenli heterojen verideki kümelenmenin, n tane p-boyutlu $x_{1,\dots,}x_n$ gözleminde her biri bilinmeyen $\pi_{1,\dots,}\pi_g$ olasılıkları ile sonlu sayıdaki g grup yoğunluklarının karmasından geldiği varsayılır[3]. Normal dağılımların karma modeli,

$$f(x_j;\theta) = \sum_{i=1}^{g} \pi_i f_i(x_j;\psi_i) \quad (1)$$

şeklinde yazılır. Burada $i = 1, \dots, g$ için π_i, $0 < \pi_i < 1$ arasında ve $\sum_{i=1}^{g} \pi_i = 1$ olacak biçimde i. küme veya grup için karma oranını göstermektedir. $j = 1, \dots, n$ için grup koşullu yoğunluk fonksiyonu $f_i(x_j;\psi_i)$, bilinmeyen parametreler vektörü ψ_i'ye bağlıdır. Bu çalışmada $f_i(x_j;\psi_i)$'nin μ_i ortalamalı Σ_i varyans-kovaryans matrisli çok değişkenli normal dağılım olduğu varsayılır. Burada ψ_i, bileşenlerin parametre vektörüdür ve $\psi_i = (\mu_i, \Sigma_i)$ şeklinde gösterilir. $f_i(x_j;\mu_i,\Sigma_i)$ çok değişkenli olasılık yoğunluk fonksiyonu,

$$f_i(x_j;\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_i|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x_j-\mu_i)^T {\Sigma_i}^{-1}(x_j-\mu_i)\right\} \quad (2)$$

denklemi ile verilir. (2)'deki eşitlikte T üst indisi, matrisin transpozunu göstermektedir. Böylece (1)'deki eşitlikte $\theta = \pi_1, \dots, \pi_g,\ \psi_1, \dots, \psi_g$ vektörü, Ω parametre uzayında çok değişkenli normal dağılımların karmasının bilinmeyen parametrelerinin tümünü temsil eden vektördür .

Çok değişkenli veri setinde çoklu kümelenme yapılarını tanımlamada değişkenlerin yapıları ve parçalanmaları veriye en uygun küme sayısını bulmada kullanılır [4]. Karma normal dağılımlarla kümeleme yapmak için kök seçim metodu verideki alt grup yapılarını tanımlamak üzere geliştirilmiştir [5]. Çok değişkenli sonlu karma modellerde değişkenlerdeki yapının belirlenmesi verideki maksimum küme sayısının belirlenmesi ile başlar. Belirlenen bu stratejide küme sayısının belirlenmesinde Genetik Algoritmalar kullanılır (GA) [6]. Normal dağılımların karmasındaki bütün parametreler, beklenti ve maksimum yapma (Expectation and Maximazation - EM) algoritmasıyla en çok olabilirlik (Maximum Likelihood) metodu kullanılarak belirlenir [7]. Veri için en iyi kümelenme modeli, Akaike bilgi kriteri (AIC) [8] ve Bayesçi bilgi kriteri (BIC) [9] gibi bilgi kriterleri optimizasyon amaçlı kullanılabilir. AIC değeri minimum olan karma dağılım modeli optimum model olarak belirlenir ancak, AIC verideki kümelenme sayısını olduğundan fazla göstermeye meyilli olduğundan başka bilgi kriterlerinede ihtiyaç vardır. Modele dayalı kümeleme yapmak için verilerdeki değişkenlerin parçalanmasına bağlı yeni bir Genetik Algoritma kullanarak karma normal dağılımlardan model sayılarıyla ilgili bir aralık elde edilir [10]. Büyük verideki ızgara yapılarda modele dayalı karma normal modellerin kümelenmesi için değişken veri segmentasyonu kullanılır [11]. Uzaktan algılama görüntü verisinin karma normal modele dayalı yarı denetimli (semi-supervised) sınıflandırılmasında Genetik Algoritmalar kullanılır [12]. Sonlu karma modellerdeki model sayılarının ve yapılarının doğru ve etkili elde edilmesinde bilgi kriterleri kullanılır [13].

Bu çalışmada homojen ve heterojen değişkenler içeren on beş değişkenli verideki kümelenmeyi karma normal dağılımları kullanarak belirlemek için yeni bir kümeleme algoritması geliştirilmiştir. Bu kümeleme algoritması ilk basamakta değişkenlerdeki heterojen yapıyı ortaya çıkarıp verileri kategorik verilere dönüştürmek için geliştirilen "değişken veri segmentasyonu" tekniğini kullanmaktadır. K-ortalamalar algoritması ile her bir değişkendeki küme yapısını belirleyip Genetik Algoritmalar ile modeller oluşturmaktadır. Modele dayalı kümeleme için karma normal dağılımlardan karma modelleri elde edip verideki en iyi kümelemeyi belirlemektedir.

Bu amaçla ikinci bölümde çok değişkenli heterojen verinin her bir değişkenindeki bölünme sayısına dayalı olarak parçalanmalar belirlenmiştir. Değişkenlerdeki parçalanmalara dayalı Genetik Algoritma kullanılarak verinin kümelenme yapısı ve kümelenme merkezlerinin sayısı hesaplanmıştır. Değişkenlerdeki parçalanmalar kullanılarak heterojen veride kümelenme yapısı için karma normal dağılımlardan elde edilen aday modeller oluşturulmuştur. Oluşturulan aday modellerdeki parametreler tahmin edilmiştir. Her bir normal dağılımların karması için log-likelihood, AIC ve BIC değerleri hesaplanmıştır. Hesaplanan ve elde edilen değerlerin sonucuna göre çok değişkenli heterojen verideki kümelenme yapısını belirleyen en iyi model seçilmiştir. Son bölümde bu çalışmayla ilgili yorumlar ve bazı sonuçlar verilmiştir.

2. Materyal ve Metot

Çok değişkenli verilerde her bir değişkendeki parçalanmalar tek değişkenli norma karma dağılımlar ve grafiksel yöntemler ile belirlenir [14]. Çok değişkenli veride her bir heterojen değişkendeki bölünme ya da parçalanma sayısının belirlenmesinde tek değişkenli karma modeller kullanılır. Bu çalışmada kullanılan metot ve prensiplerini teorik olarak ortaya koymak için simülasyon ile üretilen çok değişkenli sentetik veri seti kullanılmıştır. Sentetik veri setindeki her bir değişkene tek değişkenli karma normal model oluşturarak modelde bileşen sayısı belirlenir. Tek değişkenli normal dağılımların karması

$$f(x;\theta) = \sum_{i=1}^{g} \pi_i f_i(x;\mu_i,\sigma_i) \quad (3)$$

şeklinde gösterilir. Burada $f(x)$ tek değişkenli normal dağılımın karmalarının olasılık yoğunluk fonksiyonu, g karma dağılımdaki bileşen sayısı, π_i karma olasılık ağırlıklarını göstermektedir. Tek değişkenli karma normal dağılımların olasılık yoğunluk fonksiyonları $f_i(x;\mu_i,\sigma_i)$, ortalama vektörü μ ve standart sapması σ olmak üzere,

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}, x \in \mathbb{R} \quad (4)$$

şeklinde ifade edilir. Değişkenlerdeki parçalanma sayısını elde etmek için karma normal dağılımların log-likelihhod, AIC ve BIC gibi istatistiksel bilgi kriterleri kullanılmaktadır. Oluşturulan her bir tek değişkenli karma normal model için log-likelihood, AIC ve BIC değerleri değişkenlerdeki olasılık ağırlıkları π, ortalama vektörü μ ve varyansı σ^2 değerlerinden hesaplanır. Tek değişkenli karma modellerin parametreleri EM algoritması kullanılarak elde edilir. z tamamlanmış veride etiket vektörü olmak üzere EM kümeleme algoritmasında $\{X_1, X_2, \dots, X_n, Z_1, Z_2, \dots, Z_n\}$ verideki likelihood fonksiyonu,

$$L = f(x_1, x_2, \dots, x_n, z_1, z_2, \dots, z_n; \pi, \theta) = \prod_{i=1}^{n}\prod_{g=1}^{k}\left[\pi_i f\left(x;\theta_g\right)\right]^{z_{gi}} \quad (5)$$

şeklinde elde edilir. Burada likelihood fonksiyonundaki hesaplamaların daha kolay elde edilmesi için fonksiyonun logaritması alınarak log likelihood fonksiyonu,

$$lnL(\pi,\theta; x_1, x_2, \dots, x_n, z_1, z_2, \dots z_n) = \sum_{i=1}^{n}\sum_{g=1}^{k} z_{gi} ln\left[\pi_g f\left(x;\theta_g\right)\right] \quad (6)$$

olarak elde edilir. Buradaki amaç likelihood fonksiyonunun değerini en büyük yapan etiket vektörünü elde etmektir. EM algoritması iki adımdan oluşur, ilk adım (E) beklenti adımı ve ikinci adım (M) en büyük yapma adımıdır.

E Adımı: Bayesci bir kümeleme yaklaşımı olan EM algoritmasında z_{gi} değerlerini tahmin etmek için koşullu beklenen değer,

$$\hat{z}_{gi} = E\left(z_{gi}\middle|x;\pi,\theta\right) = \frac{\pi_g f(x;\theta_g)}{\sum_{g=1}^{k} f(x;\theta_g)} \quad (7)$$

şeklinde tahmin edilir.

M Adımı: Bu adımda olasılık ağırlıkları toplamı $\sum_{g=1}^{k}\pi_g = 1$ olduğundan dolayı log likelihood fonksiyonunu maksimize etmek için

$$lnL(\pi,\theta; x_1, x_2, \dots, x_n, z_1, z_2, \dots z_n) = \sum_{i=1}^{n}\sum_{g=1}^{k} \hat{z}_{gi} ln\left[\pi_g f\left(x;\theta_g\right)\right] \quad (8)$$

denklemi kullanılır. EM algoritması loglikelihood fonksiyonundaki parametre değerleri değişmeyene kadar denemeye devam eder ve parametreleri tahmin etmiş olur.

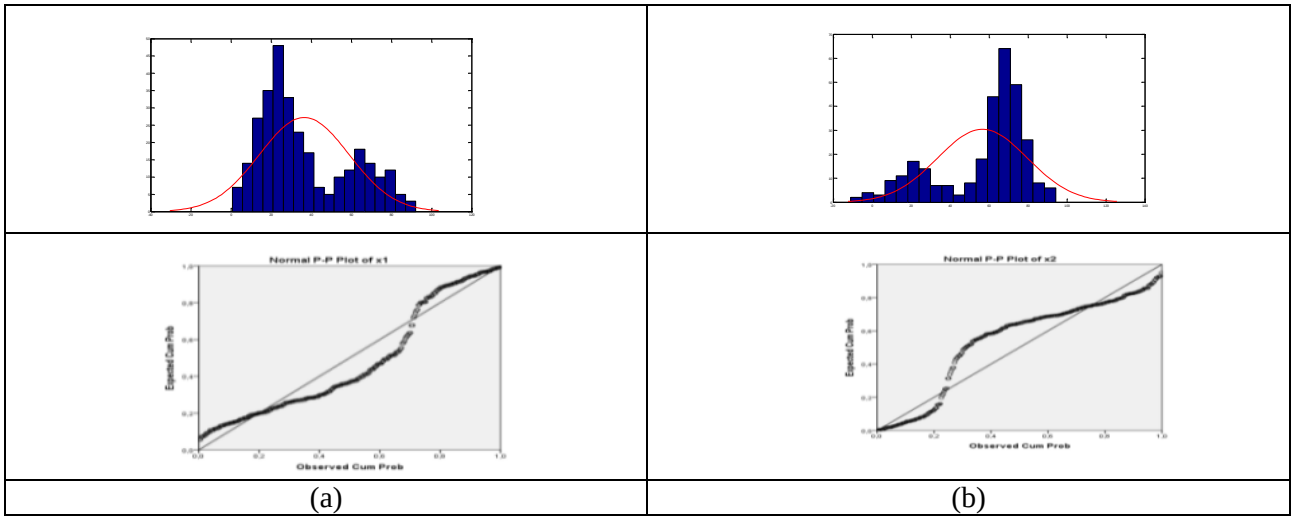
Her bir heterojen değişkendeki parçalanmayı model tabanlı belirlemek amacıyla parametreleri tahmin edilen log-likelihood fonksiyonlarından ve AIC ve BIC değerleri hesaplanır. Log-likelihood fonksiyon değerinin maksimum, AIC ve BIC değerlerinin minimum olduğu modelde parçalanma sayısı optimum olarak bulunur.

Tablo 1. Değişkenlerdeki parçalanmaların tek değişkenli karma normal modellerin Log-likelihood, AIC ve BIC değerlerine göre belirlenmesi.

Değişken No	Log-Likelihood	AIC	BIC	Parçalanma sayısı
1	-1753.4	3510.7	3510.8	2
	-1704.7	**3419.4**	**3439.4**	
	-1702.3	3420.6	3452.5	
2	-1775.8	3555.7	3563.7	2
	-1693	**3369**	**3416**	
	-1693	3402	3433.9	
3	-1485.9	**2975.9**	**2983.8**	1
	-1485.8	2981.5	3001.5	
	-1485.4	2986.7	3018.6	
4	**-1471.2**	**2948.3**	**2956.3**	1
	-1471.7	2953.4	2973.4	
	-1471.8	2959.6	2991.5	
5	-1482.6	**2969.3**	**2977.3**	1
	-1482.6	2975.3	2995.3	
	-1480	2977	3009	
6	-1489.4	**2982.8**	**2990.7**	1
	-1488.5	2986.9	3006.9	
	-1486.6	2989.2	3021.2	
7	-1495	**2993.9**	**3001.9**	1
	-1495	2999.9	3019.9	
	-1494	3003.9	3035.9	
8	-1491.4	**2986.8**	**2994.8**	1
	-1490.7	2991.3	3011.3	
	-1490.4	2996.7	3028.7	
9	-1500.6	**3005.2**	**3013.1**	1
	-1500.5	3010.9	3030.9	
	-1500.5	3017	3048.9	
10	**-1499.8**	**3003.7**	**3011.7**	1
	-1499.9	3009.8	3029.7	
	-1499.8	3015.6	3047.6	
11	-1490.6	**2985.2**	**2993.2**	1
	-1490.6	2991.2	3011.2	
	-1490.3	2996.6	3028.6	
12	-1468.8	**2941.5**	**2949.5**	1
	-1466.3	2942.6	2962.6	
	-1460	2948.1	2980	
13	-1488.2	**2980.4**	**2988.4**	1
	-1487.1	2984.2	3004.2	
	-1487	2989.9	3021.9	
14	-1478.6	**2961.2**	**2969.2**	1
	-1478.6	2967.3	2987.2	
	-1475.4	2965.1	2997	
15	-1475.3	**2954.5**	**2962.5**	1
	-1476.6	2959.3	2979.2	
	-1471.5	2959	2991.	

Çok değişkenli büyük veri setinde her bir değişkendeki parçalanma sayısını belirlemek için tek değişkenli karma normal dağılımında bileşen sayısı $k = 1,2,3$ olarak girilmiş ve üç değer arasından en iyi parçalanma sayısı bulunan istatistiksel bilgi kriterlerindeki kırılma sayısına göre belirlenmiştir. Kümeleme sayısındaki üst limit olan n sayısı grafiksel yöntemlerdeki olasılık grafiklerindeki normal doğru ile kesişme sayısına göre belirlenmektedir. On beş değişkenli büyük veride her bir değişken için elde edilen değerlere bakılarak X_1 ve X_2 değişkenlerinde uygun

parçalanma sayısı $k_1 = 2\ ve\ k_2 = 2$ olarak elde edilmiş, tablodaki diğer $X_3, X_4, \dots, X_{15}$ değişkenlerde parçalanmanın olmadığı homojen yapı gözlenmiştir.
Çok değişkenli verideki her bir değişkenin Histogram ve Normal P-P grafiklerine bakılarak değişkenlerde parçalanmaların olup olmadığına karar verilebilir.



(a) (b)

Şekil 1.On beş değişkenli veri setindeki (a) X_1 değişkeni (b) X_2 değişkeni için histogram ve P-P grafikleri.

Çok değişkenli veri setinde $X_1\ ve\ X_2$ değişkenleri için uygun parçalanmanın belirlenmesinde histogram grafiğindeki tepe (mod) sayısı ve P-P grafiğinin normallik veya $y = x$ doğrusu ile verideki gözlemlerin oluşturduğu eğrinin kesişim sayısına bakılarak parçalanma sayıları belirlenir. **Şekil 1** deki grafiklere bakılarak birinci veri setinde $k_1 = 2\ ve\ k_2 = 2$ olarak tahmin edilmiştir. X_1 değişkeni iki normal dağılımın karması, X_2 değişkeni iki normal dağılımın karmasından meydana gelmiştir.

2.1. Değişkendeki Parçalanmalara Düşen Gözlemlerin K-Ortalamalar Algoritması ile Belirlenmesi

Çok değişkenli veride değişken veri parçalanması uygulandıktan sonra elde edilen heterojen değişkenlerdeki parçalanmalara değişkendeki hangi gözlemlerin düşeceği $k-$ortalamalar algoritması ile belirlenir. Heterojen $X_1\ ve\ X_2$ değişkenlerinin parçalanmaları sırasıyla $X_{11}, X_{12}\ \ ve\ X_{21}, X_{22}$ olarak altgruplara ayrılmıştır. Veri setindeki diğer değişkenler homojen yapıda olduğundan altgrupları bulunmamaktadır. Başlangıçta her bir değişkenin parçalanma sayısı k kadar merkez sayısı belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık

$$argmin_s \sum_{i=1}^{k} \sum_{j \in S_i} \left\| x_j - \mu_i \right\|^2 \qquad (9)$$

denklemi ile hesaplanır. Parçalanmalar arası mesafenin maksimum (heterojenlik) aynı zamanda parçalanmalara düşen gözlemler arası mesafenin minimum (homojenlik) olduğu durum optimum kümelenmeyi verir $n \times 15$ tipindeki veri setinde değişken veri parçalama ve parçalanmaların bulunduğu değişkenlere k-ortalamalar algoritması uygulanarak değişkenlerdeki heterojenlik incelenmiştir. Çok değişkenli veri setinde $X_1\ ve\ X_2$ değişkenlerinde parçalanma olduğundan veri iki değişkenli $n \times 2$ formundadır. İki değişkenli veri matrisi $X = [X_1, X_2]$ şeklinde gösterilebilir. n_1 elemanlı X_1 değişkeni $X_1 = \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix}$ şeklinde gösterilir, burada X_{11} ve X_{12} sırasıyla n_{11} ve n_{12} elemanlı olup $n_1 = n_{11} + n_{12}$ olarak elde edilir. n_2 elemanlı X_2 değişkeni $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix}$ şeklinde gösterilir, burada X_{21} ve X_{22} sırasıyla n_{21} ve n_{22} elemanlı olup $n_2 = n_{21} + n_{22}$ olarak elde edilir.

Çok değişkenli veri setindeki değişkenler ve değişkenlere k-means algoritması uygulandıktan sonra değişkenlerdeki parçalamalara düşen gözlem sayıları **Tablo 2.** de verilmiştir.

Tablo 2. On beş değişkenli veri setindeki değişkenler ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

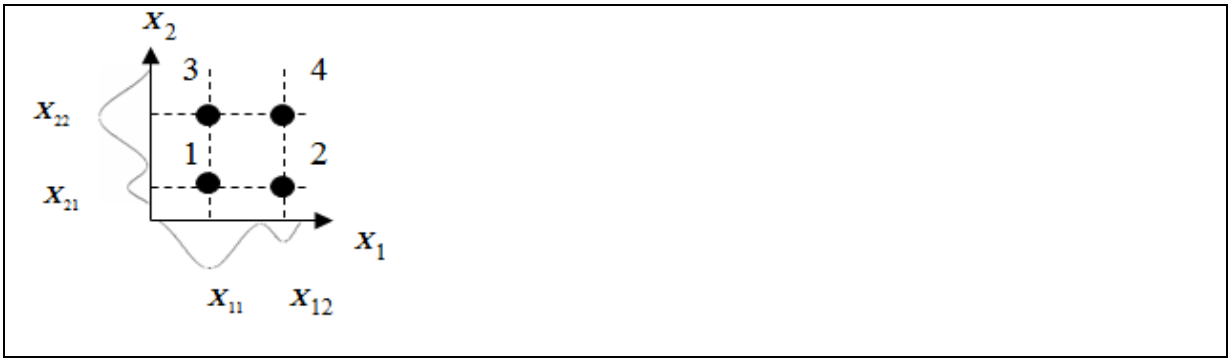
Değişken	X_1		X_2		X_3	X_4	X_5
Değişken parçaları	$X_{1.1}$	$X_{1.2}$	$X_{2.1}$	$X_{2.2}$			
Gözlem sayısı	$n_{1.1}=$ 120	$n_{1.2}=$ 280	$n_{2.1}=$ 150	$n_{2.2}=$ 250	$n_3=$ 400	$n_4=$ 400	$n_5=$ 400
Değişken	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Gözlem sayısı	$n_6=$ 400	$n_7=$ 400	$n_8=$ 400	$n_9=$ 400	$n_{10}=$ 400	$n_{11}=$ 400	$n_{12}=$ 400
Değişken	X_{13}			X_{14}		X_{15}	
Gözlem sayısı	$n_{13}=400$			$n_{14}=400$		$n_{15}=400$	

2.2. Karma Normal Modellerde Küme Merkez Sayısı ve Yerinin Belirlenmesi

Çok değişkenli X_1 *ve* X_2 değişkenin heterojen ve diğer değişkenlerin homojen olduğu veri setinde modeldeki kümelenme merkez sayıları değişkenlerdeki parçalanmalara bağlı olarak hesaplanır. Değişkenlerdeki parçalanmaların sayısı kümelenme merkez sayısını, parçalanmalara düşen gözlem değerleri ve sırasıyla, kümelenme merkezlerinin yerini, şeklini ve büyüklüğünü belirlemektedir [15]. Modeldeki değişkenlerin parçalanmalarının oluşturduğu maksimum ve minimum kümelenme merkezlerinin sayısı C_{max} *ve* C_{min}, $s = 1, \dots, p$ olmak üzere k_s değerlerine bağlı olarak,

$$C_{max} = \prod_{s=1}^{p} k_s\,,\ C_{min} = max\{k_s\} \tag{10}$$

denklemlerinden hesaplanır. Verideki parçalanmalara karşılık gelen en çok kümelenme merkezi sayısı
$C_{max} = k_1.k_2 \dots k_{15} = 2.2.1 \dots 1 = 4$ ve en az kümelenme merkez sayısı
$C_{min} = max\{k_1, k_2, \dots, k_{15}\} = max\{2,2,1, \dots, 1\} = 2$ olarak elde edilir.



Şekil 2. On beş değişkenli büyük veride X_1 ve X_2 değişkeninin alt gruplarına karşılık gelen küme merkezleri

Şekil 2. te gösterilen modelde her bir küme merkezi değişkenlerdeki alt gruplara karşılık gelir. 1. küme merkezi X_{11} *ve* X_{21} alt gruplarından, 2. küme merkezi X_{12} *ve* X_{21} alt gruplarından, 3. küme merkezi X_{11} *ve* X_{22} alt gruplarından, X_{12} *ve* X_{22} alt gruplarından meydana gelir.

2.3. Çok Değişkenli Büyük Veride Karma Normal Modellerde Toplam Model Sayısı ve Modellerin Yapısının Belirlenmesi

On beş değişkenli X_{11} *ve* X_{21} değişkenin ikiye bölündüğü ve diğer değişkenlerde parçalanmanın olmadığı durumda kümelenme merkezleri için M_{toplam} ile gösterilen oluşabilecek tüm modellerin sayısı değişkenlerdeki parçalanmalara bağlı

$$M_{Toplam} = 2^{C_{Max}} - 1 = 2^4 - 1 = 15 \tag{11}$$

olarak elde edilir. Burada çıkarılan 1 model varsayıma uymayan kümelenme merkezi bulunmayan boş modeldir.

2.4. Karma Normal Modellerde Aday Model Sayısının Hesaplanması

Çok değişkenli veride heterojen değişkenlerin alt gruplarından oluşan küme merkezlerinin sayısı ve yerine bağlı olarak model sayısı hesaplanır. Karma normal modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir küme merkezi karşılık gelir. Alt gruplara karşılık gelmeyen merkezlerin bulunduğu modeller geçerli olmayan modeller olarak hesaplamalardan çıkarılır. Geçerli modeller Şekil 2. te gösterilen merkezler üzerinden kısaca her satır ve her sütunda en az bir kümelenme merkezi bulunacak varsayımına dayanır. Varsayıma uyan modeller uygun aday modellerdir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı olarak hesaplanabilir. Alt grup bulunmayan homojen değişkenlerin küme merkezi oluşturmada ve model oluşumunda etkisi bulunmadığından karma normal modeller iki değişkenli modellerden oluşturulur.
Heterojen değişkenlerin ikiye bölündüğü durumda veride olabilecek sırasıyla minimum ve maksimum kümelenme merkez sayısı 2 ve 4 olan karma modeller için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları Tablo 3. de verilmiştir. Model sayıları arasından uygun adayların sayılarının hesaplanması zor bir problemdir. Değişken sayısının artması ve her değişkendeki parçalanma sayısına göre olabilecek küme sayılarının değişmesi ile çok parametreli bir kombinatorik probleme dönüşmektedir. Burada satır ve sütunlara düşen küme sayısına göre kombinatorik hesaptan bağıntılar elde edilmiştir.

Tablo 3. Karma normal modellerde uygun aday modeller için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları.

Merkez Sayısı	**Merkezlerin Konumu**	**Modellerin Sayısı İçin Bağıntı**	**Model Sayısı**
2 merkezli	1 1 şeklinde parçalanan model	$\binom{2}{1}\binom{1}{1}=2!=2$	2
3 merkezli	2 1 şeklinde parçalanan model	$\binom{2}{1}2!=2.2=4$	4
4 merkezli	2 2 şeklinde parçalanan model	$\binom{4}{4}=1$	1
Toplam model sayısı		$2^{2.2}-1=15$	
Toplam Uygun model sayısı		$0+0+2+4+1=7$	

Varsayım altındaki uygun aday model sayısının hesaplanmasındaki problem bir matristeki grup yapısına karşılık gelmektedir. Varsayımdaki değişkenler ve değişkenlerin parçalanmaları matrisin satır ve sütunlarına karşılık gelmektedir. n_{ij} elemanı i. satır ve j. sütundaki küme merkezini göstermek üzere, $n \times m$ tipindeki sıfırlardan oluşan bir kare matriste her satır ve her sütunda en az bir kümelenmenin olduğu başka bir ifade ile sıfırdan farklı en az bir elemanın bulunduğu matrislerin sayısı

$$f(n;k) = \sum_{i=0}^{n}(-1)^i \binom{n}{i} \sum_{j=0}^{m}(-1)^j \binom{(n-i)(m-j)}{k}$$
$$= \sum_{i,j=0}^{n,m}(-1)^{i+j} \binom{n}{i}\binom{m}{j}\binom{(n-i)(m-j)}{k} \quad (12)$$

şeklinde hesaplanır [16].
Burada n_r değişkenlerdeki veya matrisin boyutlarına karşılık gelen parçalanmaları, i_r parçalanmalardaki küme sayısını ve k karma modeldeki küme sayısını göstermektedir.
Bu yedi uygun duruma karşılık gelen kümelenme yapılarının modelleri Tablo 4. te verilmiştir.

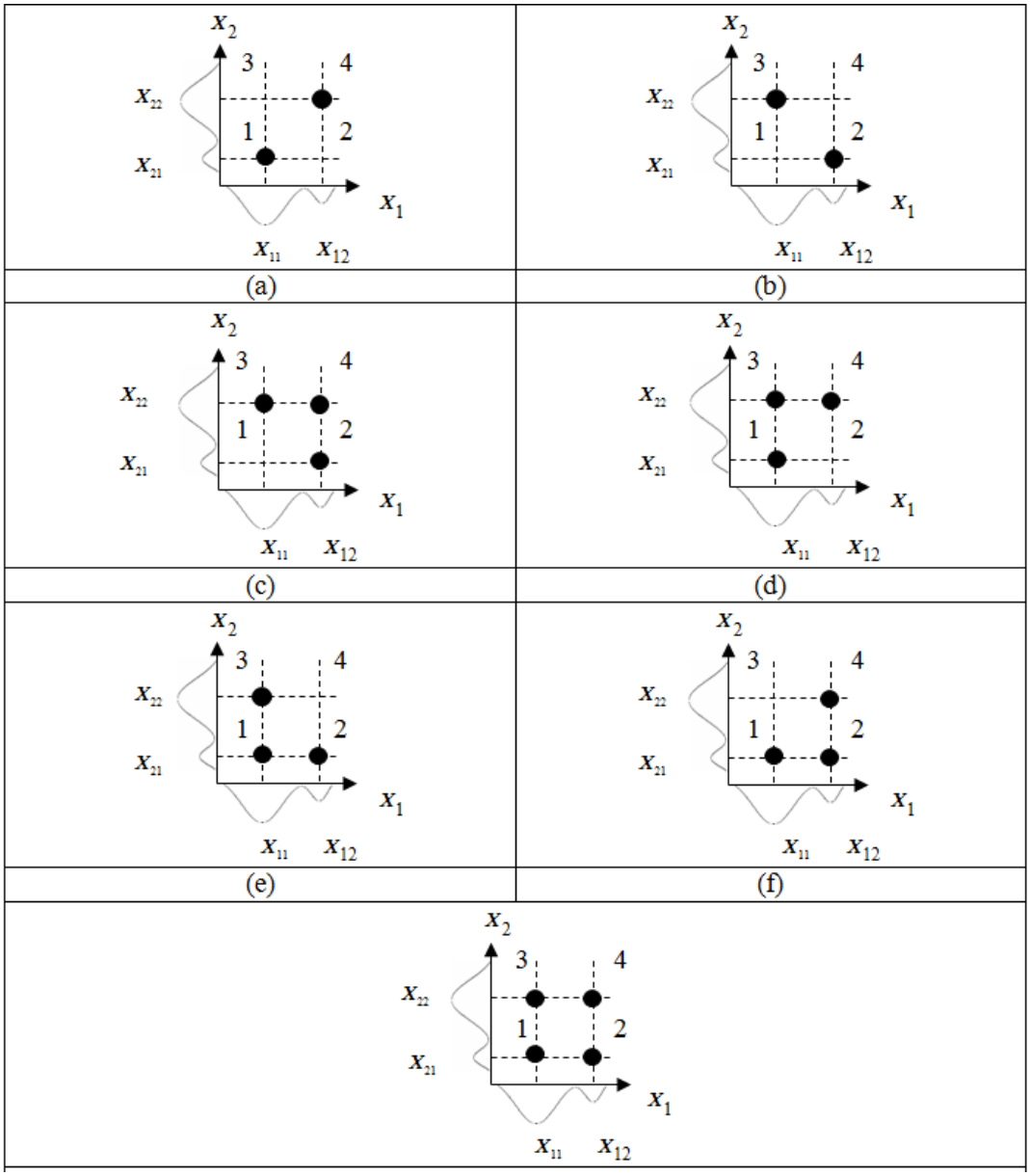
Tablo 4.Çok değişkenli X_1 ve X_2 değişkenin ikiye bölündüğü ve diğer değişkenlerde parçalanmanın olmadığı karma normal modeller arasından varsayıma uyan uygun aday modellerin sayısı, genetik algoritmaya karşılık gelen küme yapılarının dizi gösterimleri.

Kümelenme Merkez Sayısı	Durum Sayısı	Uygun Durum Sayısı	Uygun Durumlara Karşılık Gelen Modellerin Dizi Gösterimleri
			123456…15
Bir Küme Merkezli Modeller	4	-	-
İki Küme Merkezli Modeller	6	2	100100000000000
			011000000000000
Üç Küme Merkezli Modeller	4	4	111000000000000
			110100000000000
			101100000000000
			011100000000000
Dört Küme Merkezli Modeller	1	1	111100000000000

Çok değişkenli büyük veride değişkenlerdeki heterojenlik, her bir değişkene tek değişkenli karma normal model uygulanarak değişken veri parçalama metodu ile ortaya çıkarılmıştır. On beş değişkenli veri setinde Tablo 4. teki değerlere bakılarak X_1 ve X_2 değişkenlerinde $k_1 = 2$ ve $k_2 = 2$ parçalanma belirlenmiş diğer $X_3, X_4, \dots, X_{15}$ değişkenlerinde parçalanmanın bulunmadığı yani homojenlik tespit edilmiştir. Veri setindeki homojen değişkenler model oluşturmada etkisi olmadığından elenmiştir. Veri setindeki model oluşturmada etkisi bulunmayan değişkenlerin elenmesi değişken seçimi (variable selection) olarak adlandırılır. Çok değişkenli veri setinde değişken seçimi yapılınca boyut indirgeme (dimension reduction) işlemi yapılmıştır. Boyut indirgenince iki değişkenli veri setinde her bir değişkenin ikiye parçalandığı yapı oluşmuştur.

2.5. Karma Normal Modellerde Uygun Aday Modellerin Oluşturulması ve Parametre Tahminleri

İki değişkenli karma normal dağılım modelleri için bileşen ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri verideki heterojen değişkenlerdeki parçalanmalar kullanılarak örnekleme dayalı tahmin edilmektedir. Karma modeldeki her bir merkezin olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisi modeldeki olasılık yoğunluk fonksiyonunu elde etmek için kullanılır. Tablo 4. te uygun aday modellerin genetik algoritma için dizi gösterimleri modeldeki merkezlerin numaralarına göre verilmiştir. Genetik algoritmayı model üzerinde uygularken modeli temsil eden dizdeki (DNA Sequence) “0” ve “1” elemanları (gen) modeldeki karşılık gelen merkezde yığılmanın (kümelenmenin) olup olmadığını göstermektedir. Eğer uygun aday modelde herhangi bir merkezde kümelenme varsa “1” yoksa “0” ile gösterilmiştir. Veri setindeki değişkenlerin her birisi matris gösteriminde satır ve sütunlara karşılık geldiğinden, iki boyutlu düzlemde uygun durumlar ve bu durumların merkezlerinin yerlerinin anlatıldığı modeller elde edilmiştir. İki değişkenli ve her değişkenin ikiye parçalandığı veri setinde oluşabilecek uygun aday modellerin dizi gösteriminden elde edilen modellerin ızgara yapısı Şekil 3. te verilmiştir.



Şekil 3. Çok değişkenli veri setinde (a) ve (b) iki merkezli karma normal, (c), (d),(e) ve (f) üç merkezli karma normal modeller ve (g) dört kümelenme merkezli uygun aday karma normal modellerin düzlemsel grafiklerini göstermektedir.

Karma normal modellerdeki k küme merkez sayısını $i, j = 1{,}2$ olmak üzere, Uygun aday modellerin ortalama vektörü μ_k, varyans-kovaryans matrisi Σ_k ve korelasyok katsayısı ρ_k sırasıyla $\mu_k = \begin{bmatrix}\mu_{1i}\\ \mu_{1_j}\end{bmatrix}$, $\Sigma_k = \begin{bmatrix}\sigma_{1i}^2 & \rho_k\sigma_{1i}\sigma_{2j}\\ \rho_k\sigma_{2i}\sigma_{1j} & \sigma_{2j}^2\end{bmatrix}$ ve $\rho_k = Corr(X_{1i}X_{2j}) = \frac{\sigma_{1i2j}}{\sigma_{1i}\sigma_{2j}}$
olacak şekilde elde edilir. Bu durumda , k. küme merkezindeki veriler, , $N(x;\mu_k,\Sigma_k)$ çok değişkenli normal dağılıma sahip olur.

2.6. Genetik Algoritma ile Karma Normal Model Oluşturulması

Çok değişkenli veri setindeki iki değişkenin her birinin ikiye bölünmesi durumunda oluşacak kümelenme merkezleri Şekil 2. te gösterilmiştir. Tablo 4. te her bir veri setindeki parçalanmalara karşılık gelen ve varsayımlara uyan uygun aday modellerin dizi gösterimi verilmiştir.
Toplam 7 uygun aday model içerisinden iki, üç ve dört bileşenli karma normal modellere karşılık gelen modeller $u = 1, \dots, 4$ ve $i = 1, \dots, 4$ için, $0 < \pi_i < 1$ ve $\sum_{i=1}^{4}\pi_i = 1$ olmak üzere normal bileşen yoğunluk fonksiyonları

$$f\left(x;\mu^{(u)},\Sigma^{(u)}\right) = \sum_{i=1}^{4}\pi_i f_i\left(x;\mu^{(u)},\Sigma^{(u)}\right) \quad (13)$$

olmak üzere, olasılık ağırlıkları $\pi_i{}^{(u)} = \frac{\pi_i}{\sum_{i=1}^{4}\pi_i}$, $x = 1, \dots, 4$ ve $y = 1, \dots, 4$ olmak üzere ortalama vektörleri $\mu_i{}^{(u)} = \begin{bmatrix} \mu_{1x}{}^{(u)} \\ \mu_{2y}{}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_k{}^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}{}^{(u)}\sigma_{1x}{}^{(u)}\sigma_{2y}{}^{(u)} \\ \rho_{2y,1x}{}^{(u)}\sigma_{2y}{}^{(u)}\sigma_{1x}{}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir.

3. Bulgular

Modele dayalı kümeleme yapmak için on beş değişkenli veride homojen değişkenler değişken seçimi ile elenip veride boyut indirgeme işleminden sonra iki değişkenli ve her bir değişkenin ikiye parçalandığı veri setindeki uygun aday modellerin karma normal modelleri genetik algoritma ile belirlenmiştir. Karma normal modeller arasından en iyi modelin seçimi uygun aday modellerin DNA dizilimindeki (1001) her bir gen ("0" ve "1") elemanlarına atanan ağırlıklar ve parametrelere göre istatistiksel öğrenme yöntemleri ile log-likelihood, AIC ve BIC gibi bilgi kriterleri yardımı ile belirlenir. Log-likelihood fonksiyonu

$$logL(\Psi) = \sum_{j=1}^{n} log\left(\sum_{i=1}^{g}\pi_i f_i(x_j;\theta_i)\right) \quad (14)$$

olarak elde edilir.

Çok değişkenli karma normal modellerin log-likelihood fonksiyonuna bağlı olarak Bayesci bilgi kriteri (BIC)

$$BIC = -2logL(\Psi) + dlogn \quad (15)$$

olarak elde edilir. Burada n olasılık yoğunluk fonksiyonundaki gözlem sayısını, K bileşen veya grup sayısı ve p değişken sayısına bağlı olarak d modeldeki bağımsız parametre sayısını

$$d = (K-1) + (Kp) + \left(Kp\,\frac{(p+1)}{2}\right) \quad (16)$$

elde edilir. Akaike bilgi kriteri de (AIC) log-likelihood fonksiyonuna bağlı olarak

$$AIC = -2logL(\Psi) + 2d \quad (17)$$

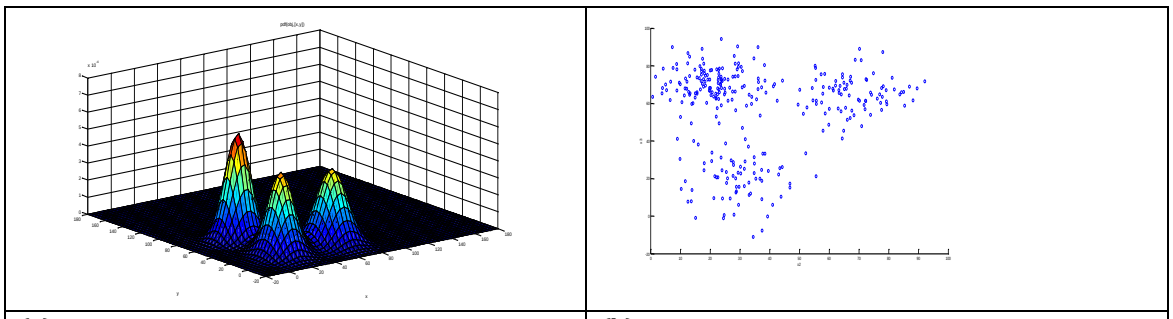
şeklinde elde edilir.

On beş değişkenli büyük verilerin parçalanması ile eide edilen segmentasyona dayalı değişken seçimi (variable selection) yapılarak homojen yapıdaki değişkenler çıkarıldıktan sonra iki değişkendeki ikişer parçalanmadan kaynaklanan küme merkezlerine göre uygun aday modellerin log-likelihood, AIC ve BIC değerlerinin hesaplandığı değerler Tablo 5. te verilmiştir.

Tablo 5. Genetik algoritma ile belirlenen İki değişkenli karma normal dağılımların modele dayalı kümelenmesi için Log L, AIC ve BIC değerleri.

Model	$\text{Log } L(\Psi)$	$AIC = -2\ln L(\Psi) + 2d$	$BIC = -2\ln L(\Psi) + d\log n$	d
1.	-3979	7980	7958	11
2.	-4544.6	9111.1	9089.1	11
3.	-4091.8	8217.6	8183.6	17
4.	-4103.8	8241.5	8207.5	17
5.	**-3466.5**	**6967**	**6933**	**17**
6.	-4050.2	8134.5	8100.5	17
7.	-3550.9	7147.9	7101.9	23

Modele dayalı kümeleme için çok değişkenli karma normal modeller arasından en iyi modelin seçimi modellerin log-l, AIC ve BIC değerlerine dayalı olarak belirlenmektedir. Çok değişkenli veri setinde karma normal modellerden uygun aday modellerin log-l, AIC ve BIC değerleri hesaplanmış ve Tablo 5. de verilmiştir. Uygun aday modeller arasından modele dayalı kümelemede en iyi model, log-likelihood değeri en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir. Elde edilen değerlere göre veri setindeki değişkenlerin parçalanmalarına düşen parametrelerinden elde edilen üç kümelenme merkezli ve Tablo 4'te verilen **101100000000000** gen dizilimine sahip karma normal Model (GMM) en iyi modeldir.



(a) (b)

Şekil 4. On beş değişkenli büyük verideki $\mathbf{X_1}$ **ve** $\mathbf{X_2}$ değişkenlerindeki alt gruplardan oluşan modeller arasından veri yapısına uyan en iyi modelin (a) yoğunluk fonksiyonunun yüzey grafiği (b) saçılım grafiği.

4. Tartışma ve Sonuç

Çok değişkenli büyük veride değişkenlerin parçalanmasına dayalı karma modellerin elde edilmesi ve bu modeller arasından Genetik Algoritmalarla (GA) uygun modellerin belirlenmesi modele dayalı karma normal modeller ile belirlenmiştir. Aşamalı olarak gerçekleşen kümeleme algoritmasında veri yapısına uygun değişken seçimi ve bu heterojen değişkenler kategorilere dönüştürülmüştür. Karma modeller elde edildikten sonra modeldeki parametreler verideki olasılık ağırlıkları, ortalama vektörü ve kovaryans matrisleri kullanılmıştır. Böylece kümeleme algoritması için bilgi karmaşıklığı oluşmamış ve modeller hızlı aynı zamanda eksiksiz olarak elde edilmiştir. Elde edilen karma normal modeller arasından veriye en uygun kümeleme yapısı optimizasyon ile Değişken Veri Segmentasyonuna dayalı Karma Normal Model (GMM) kümeleme ile elde edilmiştir.

Kaynakça

[1] Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 97, 611-631.

[2] Fraley, C. and Raftery, A. E., 1998. How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. The Computer Journal, 41, 578-588.

[3] McLachlan, G. J. and Chang, S. U. (2004). Mixture Modelling for Cluster Analysis. Statistical Methods in Medical Research 13, 347-361.

[4] Galimberti, G. and Soffritti, G. (2007). Model-based methods to identify multiple cluster structures in a data set. Computational Statistics and Data Analysis. doi 10.1016/j.csda.2007.02.019.

[5] Seo, B. and Kim, D. (2012). Root selection in normal mixture models. Computational Statistics and Data Analysis. 56, 2454-2470.

[6] Nguyen, T. T., Liew, A. W. C., Tran, M. T., & Nguyen, M. P. (2014, August). Combining multi classifiers based on a genetic algorithm–a gaussian mixture model framework. In International Conference on Intelligent Computing (pp. 56-67). Springer, Cham.

[7] McLachlan, G. J. and Krishnan, T. (1997). The EM Algorithm and Extensions. New York, Wiley.

[8] Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6): 716–723.

[9] Schwarz, G., 1978. Estimating the dimension of a model, Ann. Statist. 6 pp. 461–464.

[10] Servi, T. and Erol, H., 2007. On Total Number Of Candidate Component Cluster Centers And Total Number of Candidate Mixture Models In Model Based Clustering. Selçuk Journal of Applied Mathematics Vol.8. No.2. pp. 57 – 69.

[11] Erol, H. Gogebakan, M. Erol, R. (2017) Grid Structures and Orientations Of Clusters Using Discretization Of Variables In Big Data. Proceedings of International Conference on Engineering, Technology, and Applied Science ICETA 2017, ISSN 2411-9318, pp. 16-31.

[12] Gogebakan, M., & Erol, H. (2018). A New Semi-supervised Classification Method Based on Mixture Model Clustering for Classification of Multispectral Data. Journal of the Indian Society of Remote Sensing, 46(8), 1323-1331.

[13] Akogul, S., & Erisoglu, M. (2017). An Approach for Determining the Number of Clusters in a Model Based Cluster Analysis. Entropy, 19(9), 452–0

[14] Gogebakan, M., & Erol, H. (2019). Mixture Model Clustering Using Variable Data Segmentation and Model Selection: A Case Study of Genetic Algorithm, Mathematics Letters. Vol. 5, No. 2, 2019, pp. 23-32. doi: 10.11648/j.ml.20190502.12

[15] Erol, H., 2013. A model selection algorithm for mixture model clustering of heterogeneous multivariate data. In Innovations in Intelligent Systems and Applications. 2013 IEEE International Symposium on

Innovations in Intelligent Systems and Applications, At Albena, Bulgaria. (pp. 1-7). DOI: 10.1109/INISTA.2013.6577617

[16] Cheballah, H., Giraudo, S., & Maurice, R., 2015. Hopf algebra structure on packed square matrices. Journal of Combinatorial Theory, Series A, 133, 139-182.